# Essentials of Data Science with R Software-1

**Professor Shalabh**

**Department of Mathematics and Statistics**

**Indian Institute of Technology Kanpur**

**Lecture 67**

## One Sample Test for Mean with Unknown Variance

Hello friends. Welcome to the course Essentials of Data Science with R Software-1 in which we are trying to understand the basic fundamentals of probability theory and statistical inference that are useful for data science. So, now you can recall that in the last lecture, we started a discussion and we developed a test statistic that can be used for testing the hypothesis related to the mean of a normal population.

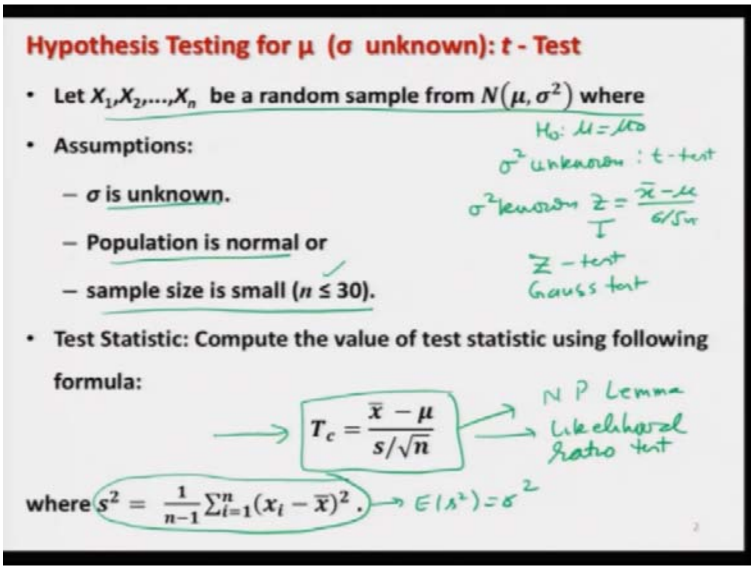
And in that case, we had assumed that $\sigma^2$ is known. Now, the next question is what to do if a $\sigma^2$ is unknown? And actually, in practice if you try to see $\sigma^2$ is unknown to us. So, the most simple option is that we can estimate it on the basis of a given sample of data and we can replace it in the place of $\sigma^2$ and try to develop a statistic which can be used for testing of hypothesis related to the mean.

Well in the last lecture I had given you the developments in detail, means how the critical reasons are located, how the decisions are taken on the basis of critical region and the nature of hypothesis whether it is left tail test, right tail test or two tail test. So, in this lecture I will not be giving you that much detail because the steps, procedure everything is the same. The only thing is this you need to know basically the statistic and how to get it done in the R software. But before going into the lecture one thing, I would like to address you, that is very important to understand.

There are two topics which you have done. One is estimation of the parameters and another is testing of hypothesis that we are doing. Remember one thing the topics which we have done in the estimation of parameter, they are helpful in finding out the good value of the unknown parameter that is existing in the population. Now, in the case of testing of hypotheses, we are only trying to compare. Suppose the null hypothesis that $H_0 : \mu = 8$ is not accepted, then the question comes well $\mu = 8$ is not true that means $\mu$ is not equal to 8 but what is the correct value of mu. That cannot be answered from the tools of test of hypothesis.

These tools can only help you in making a comparison. Comparison with a hypothetical value or comparison between two samples or comparison between more than two samples. So, this is very important point that you have to keep in mind. So, let us now begin our lecture and I hope that you had a revision of the last lecture because it is very important to understand the topics in this lecture and in the next lecture.

(Refer Slide Time: 03:27)



**Hypothesis Testing for μ (σ unknown): *t* - Test**

- Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$ where
- Assumptions:
  - $\sigma$ is unknown.
  - Population is normal or
  - sample size is small ($n \le 30$).
- Test Statistic: Compute the value of test statistic using following formula:

$$\longrightarrow \boxed{T_c = \frac{\bar{x} - \mu}{s/\sqrt{n}}}$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. $\rightarrow E(s^2) = \sigma^2$

$H_0: \mu = \mu_0$

$\sigma^2$ unknown : t-test

$\sigma^2$ known $Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

Z-test

Gauss test

N P Lemma

Likelihood ratio test

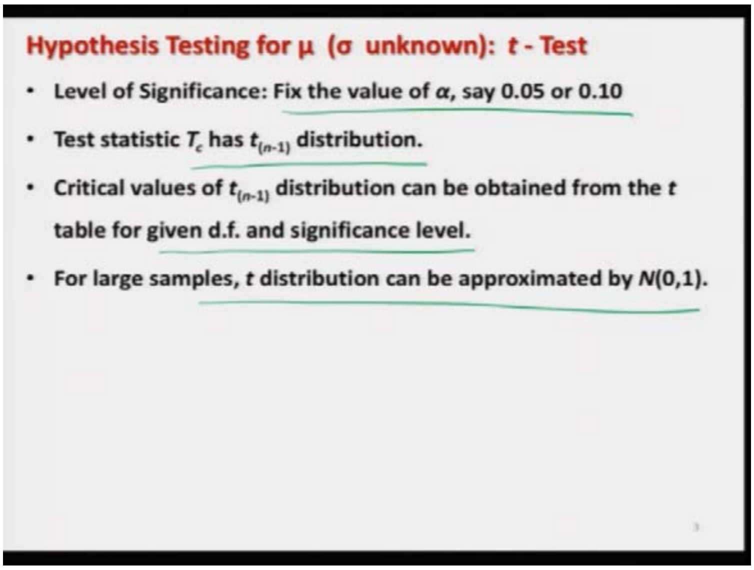
So, let us begin our lecture. So, now we are going to understand the test for $H_0$ μ is equal to $\mu_0$ and where we are trying to assume that $\sigma^2$ is unknown. So, this test essentially is a in a common language this is called as a t test because you will see that this is based on the t statistics and similarly when $\sigma^2$ was known to us, then we had used the stat z=x bar minus μ upon σ by root n.

So, since this is based on the Z statistic, so many times people in common language they call it as a Z test or Gauss test. So, these are the different names so you do not get confused, it means if you are trying to listen these names instead of the test procedures. So, in this case we assume that let $X_1$, $X_2$,.., $X_n$ be a random sample of size n from normal μ $\sigma^2$ and we assume that σ is unknown population is normal and sample size is small. What is the meaning of small? n less than equal to 30. Now, I do not need to explain you why this condition is coming because you can recall that we had discussed that when the degrees of freedom of the t distributions are 30 or more than 30, then the normal distribution become the same means their structure will become the same, their probabilities become the same.

Now, in this case, we use the test statistics which is given by here like this $T_c = \frac{\bar{x} - \mu}{s/\sqrt{n}}$. So, what we try to do because $\sigma^2$ is unknown so we try to estimate the $\sigma^2$ by $s^2$ and this is $s^2$ is an unbiased estimator of the $\sigma^2$. So, we try to replace $\sigma^2$ by $s^2$ and we have this statistic, now the question comes how this statistic is coming.

Well, it is not coming from sky but this has been derived using the Neyman Pearson lemma which is briefly called as NP lemma or this can also be derived using the likelihood ratio test. So, do not think that it is coming from the sky or well if you wish you can just find out the likelihood function under $H_0$ under $H_1$ and then try to take the ratio, try to solve them try to find out the critical reason and you will come to know that this is the statistics. Now, what we have to do? We have to get the sample and we have to compute the value of this Tc on the basis of given sample of data.

(Refer Slide Time: 06:04)



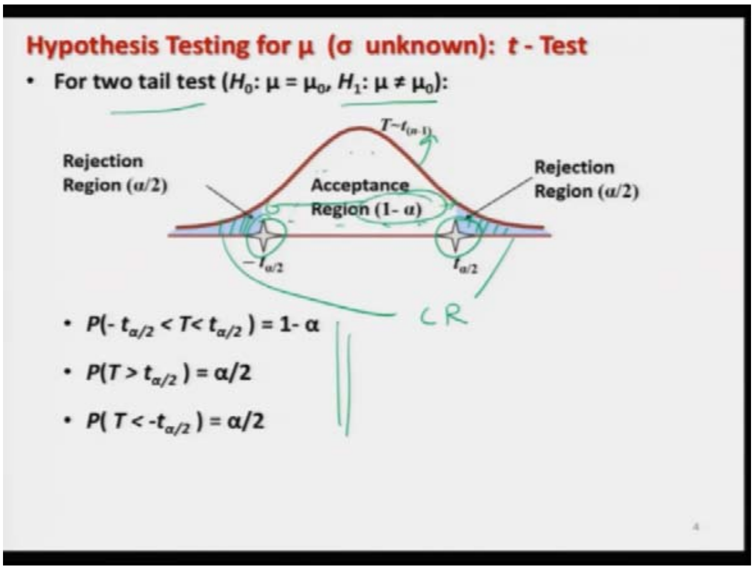
Hypothesis Testing for μ (σ unknown): *t* - Test

- Level of Significance: Fix the value of $\alpha$, say 0.05 or 0.10
- Test statistic $T_c$ has $t_{(n-1)}$ distribution.
- Critical values of $t_{(n-1)}$ distribution can be obtained from the *t* table for given d.f. and significance level.
- For large samples, *t* distribution can be approximated by *N*(0,1).

Now, how to compare it? So, for the comparison, we can fix the value of  that is the level of significance to be say 5 percent, 1 percent or anything whatever you want depending on the requirement and then we try to find out the probability distribution of this statistic $T_c$ and then we can find that $T_c$ has got a t distribution with n – 1 degrees of freedom and you remember that when we had done the t distribution also then also we had discussed this topic that the distribution of such a statistics is t distribution.

Now, the critical values of this t distribution can be obtained from the t tables for a given degrees of freedom and value of  and in case if degrees of freedoms are more than a 30, then instead of finding out that t values one can use the critical values from the standard normal distribution. There is no issue and in case if you are trying to do it on the software, then there is no issue you can simply compute the p-value.

(Refer Slide Time: 07:05)



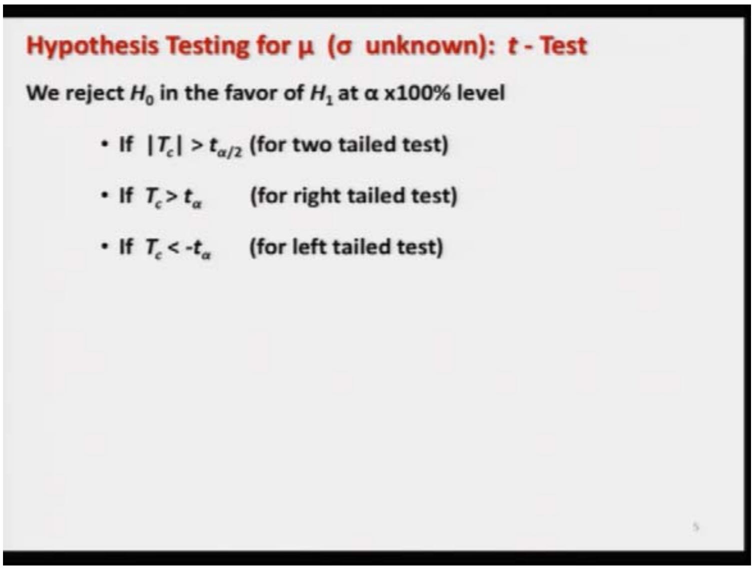
Hypothesis Testing for μ (σ unknown): *t* - Test

- For two tail test ($H_0$: $\mu = \mu_0$, $H_1$: $\mu \neq \mu_0$):

- $P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha$
- $P(T > t_{\alpha/2}) = \alpha/2$
- $P(T < -t_{\alpha/2}) = \alpha/2$

Now, as we did earlier in case if you want to test our hypothesis like $H_1$: $\mu \neq \mu_0$ that means we have a two tailed test then the critical reasons are going to be on the both the sides of the distribution. This curve is the distribution of t distribution with n – 1 degrees of freedom. So, this region is going to be α/2, this region is going to be α/2 and these are the critical values at α/2 level of significant that can be obtained from the table or from the software. And so, these are the things so you can see here this is the region here which is 1 – α. This region exactly on the same way as we have done earlier.

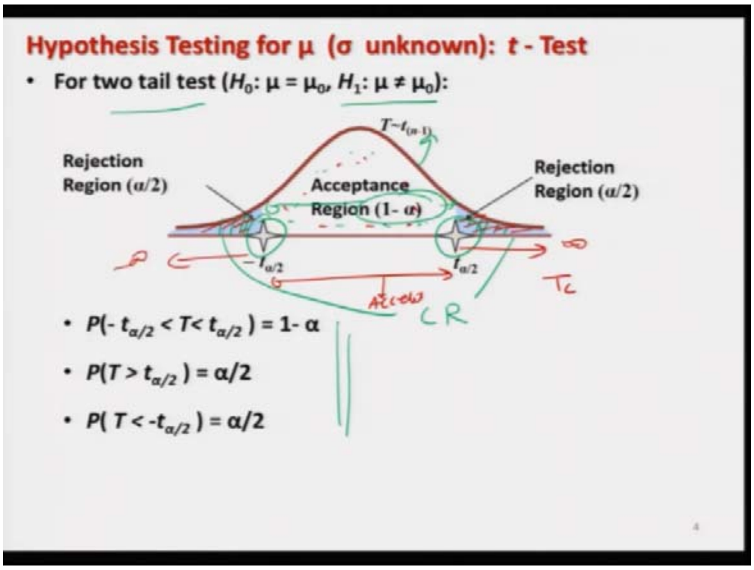
(Refer Slide Time: 07:48)



## Hypothesis Testing for μ (σ unknown): *t* - Test

We reject $H_0$ in the favor of $H_1$ at α x100% level

- If $|T_c| > t_{\alpha/2}$ (for two tailed test)
- If $T_c > t_\alpha$ (for right tailed test)
- If $T_c < -t_\alpha$ (for left tailed test)



## Hypothesis Testing for μ (σ unknown): *t* - Test

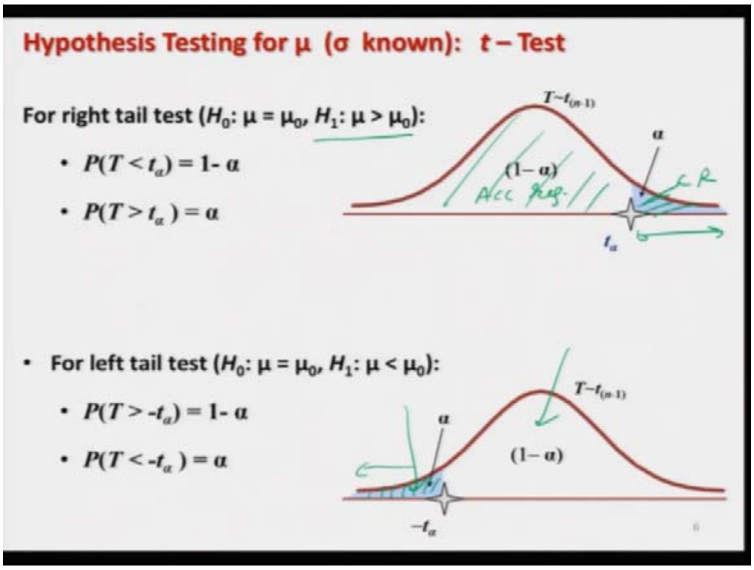
- For two tail test ($H_0$: $\mu = \mu_0$, $H_1$: $\mu \neq \mu_0$):

- $P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha$
- $P(T > t_{\alpha/2}) = \alpha/2$
- $P(T < -t_{\alpha/2}) = \alpha/2$

And in this case what are you going to do? If you want to accept or reject a hypothesis in a very simple language you are simply trying to do the same thing that we did earlier that you try to compute c and then try to see where this $T_c$ is going to be in this graph. Here in the critical region or here in the acceptance region.

So, wherever this $T_c$ is a there you try to accept or reject the hypothesis. For example, if $T_c$ is lying between $-t_{\alpha/2}$ and $t_{\alpha/2}$ then we accept it and if where $T_c$ is greater than $t_{\alpha/2}$ or it is smaller than $-t_{\alpha/2}$ then we are going to reject it. As simple as that and this is what precisely I have written here. So, if you try to understand it from the graphic point of view, I personally feel that it becomes very easy to understand.

(Refer Slide Time: 08:39)



Hypothesis Testing for μ (σ known): *t* – Test

For right tail test ($H_0$: $\mu = \mu_0$, $H_1$: $\mu > \mu_0$):

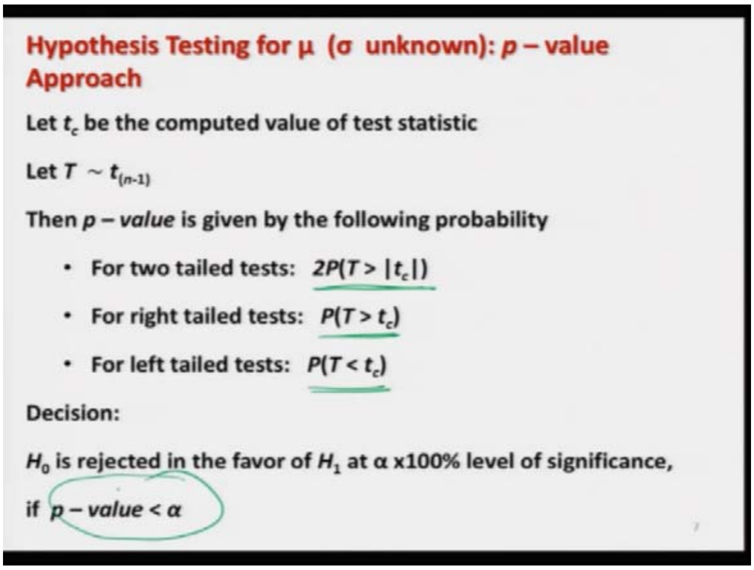
- $P(T < t_\alpha) = 1 - \alpha$
- $P(T > t_\alpha) = \alpha$

- For left tail test ($H_0$: $\mu = \mu_0$, $H_1$: $\mu < \mu_0$):
  - $P(T > -t_\alpha) = 1 - \alpha$
  - $P(T < -t_\alpha) = \alpha$

Now, similarly in the case of $H_1 : \mu > \mu_0$that is a right tail test the critical region is going to be on the right-hand side, compute the value of $T_c$ and try to see where it is lying. So, this is your here acceptance region and this is your here critical region. Try to see where the capital Tc value or the calculated value of the t statistics is going to lie if this is lying here in the acceptance region except the hypothesis, if it is lying in the critical region try to reject the hypothesis.

And similarly for the left tail test also, the critical reason is going to be on the left-hand side. Just try to find out the calculated value of the statistics and try to see whether it is lying in the critical region or in the acceptance region and then based on that you try to accept or reject the null hypothesis. Whenever we are talking of the accepting of or rejecting the hypothesis and I am not saying any name like alternative or null, then the null hypothesis is the default. We say accept the hypothesis reject the hypothesis that means we are referring to $H_0$.

(Refer Slide Time: 09:44)



**Hypothesis Testing for μ (σ unknown): $p$ – value Approach**

Let $t_c$ be the computed value of test statistic

Let $T \sim t_{(n-1)}$

Then $p$ – *value* is given by the following probability

- For two tailed tests: $2P(T > |t_c|)$
- For right tailed tests: $P(T > t_c)$
- For left tailed tests: $P(T < t_c)$

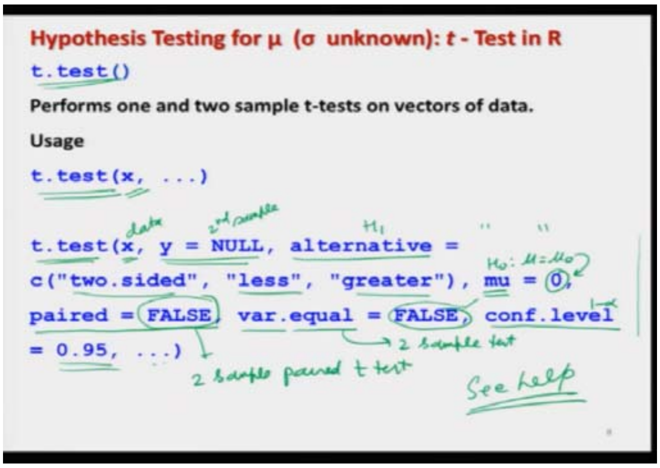
Decision:

$H_0$ is rejected in the favor of $H_1$ at α x100% level of significance,

if $p - value < \alpha$

That is our understanding. So, now in case if you want to do the test of hypothesis on the basis of p-value, so you do not have to do anything the software will compute it and then you have to just look into the value of p and then you have to take a call whether you are going to accept or reject the hypothesis. So, in this case for two tail tests the p-value is defined like this, for right tail test like this and for left tail test this like this one. So, that that is simply straight forward and $H_0$ is rejected at a 100 α percent level of significance if p-value is smaller than α. This is what you have to just keep in mind, as simple as that.

(Refer Slide Time: 10:28)



**Hypothesis Testing for μ (σ unknown): $t$ - Test in R**

`t.test()`

Performs one and two sample t-tests on vectors of data.

Usage

```
t.test(x, ...)

t.test(x, y = NULL, alternative =
c("two.sided", "less", "greater"), mu = 0,
paired = FALSE, var.equal = FALSE, conf.level
= 0.95, ...)
```

data
2nd sample
H1
H0: μ=μ0
1-α
2 sample test
2 sample paired t test
See help

Now, the next question comes how are you going to do or conducts such a test of hypothesis in the R software. So, for that this procedure is built in in the base package of R, you do not need to employ or install any additional package for this, as this was the case in the case of when $\sigma^2$ is known, so the command here is key dot test and this performs one sample, two sample t test for different type of conditions.

So, here I am trying to first give you the details but we are going to use here the details only for this test and remaining details you will see we will be using in the next lecture when we try to consider the test of hypothesis for two samples. So, now the command here is simply here t test. So, the more simple is this you simply try to write down here t dot t-e-s-t all in small αbets, lower case αbets and give here the data vector x but in general, if you try to look into the help this will give you a this type of command and where I am trying to explain you now each and everything and later on we will use it.

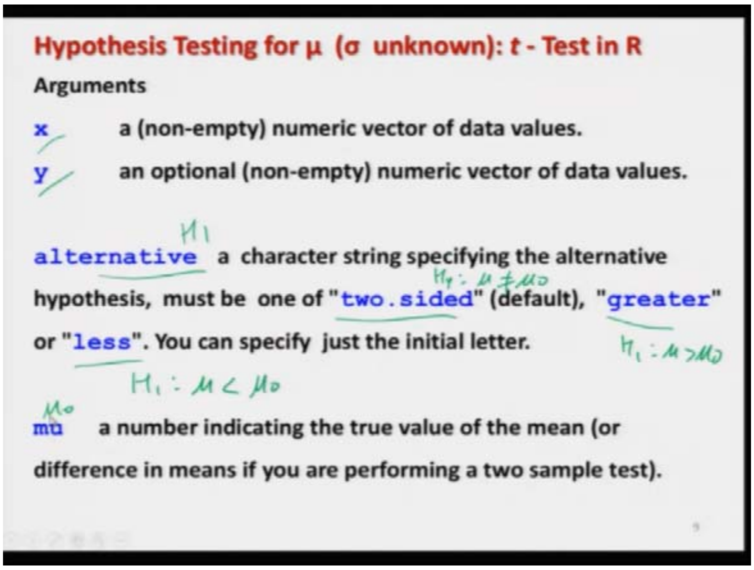
So, command here is t dot test, x is the data vector, y is the data vector for the second sample but at this moment we are not considering the second sample, we are doing only the one sample t test. So, that is why I am writing here as a null and you double it in the capital letters and then I have to specify the alternative hypothesis $H_1$. So, for that I have three choices, two sided or less or greater. So, I have to choose any one of them and I simply have to write down within the double quotes the value like as two dots guided or less or greater depending on the requirement. $\mu$ here it is the value of here $H_0 : \mu = \mu_0$.

So, this is the value of here $\mu_0$. So, by default, it is taken to be 0 but you can give here any value. Now, the next option here is p-a-i-r-e-d paired. So, paired is related to when we are trying to do the two sample paired t test. That means when the observations are paired. So, that is why this is a logical variable, it can take value TRUE or FALSE. So, whenever you are trying to conduct the two sample pair t test, then this value will simply become true that is all and now there is another option here v-a-r dot e-q-u-a-l that is variance equal. Well, this is also valid for two sample test that we are going to discuss in the next lecture.

So, when you are trying to take the two-sample test, then you have two samples which are coming from normal population and there can be two situations whether the variances of the normal population are the same or they are different. So, based on that you have to specify here whether the variances are equal or not in terms of true or false. So, if here in the present case since I am dealing only with the one sample test, so I will put it here as a false. Now, this is here confidence level that means the value of $1 - \alpha$. Remember one thing this is the value of here $1 - \alpha$. So, this is given by c-o-n-f dot l-e-v-e-l and here you have to specify the value so for example here I am writing 0.95 and after that there are many other option.

I will simply ask you to look into the help, that will give you more details. So, now today I have explained you the use of this command and the same command we are going to use when we are trying to consider the two sample test. But there I will simply refer you that there we had discussed and you have to remember these things.

(Refer Slide Time: 14:28)



Hypothesis Testing for μ (σ unknown): *t* - Test in R

Arguments

x a (non-empty) numeric vector of data values.

y an optional (non-empty) numeric vector of data values.

alternative a character string specifying the alternative hypothesis, must be one of "two.sided" (default), "greater" or "less". You can specify just the initial letter.

$H_1$

$H_1: \mu \neq \mu_0$

$H_1: \mu > \mu_0$

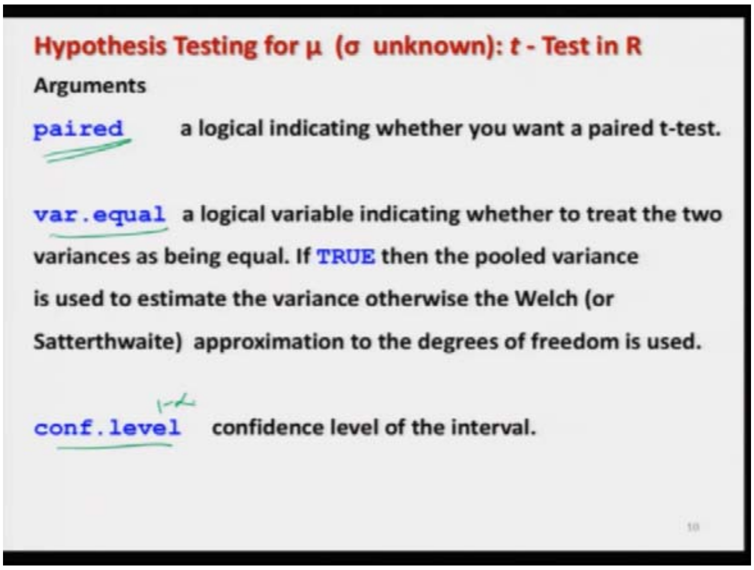
$H_1: \mu < \mu_0$

$\mu_0$

mu a number indicating the true value of the mean (or difference in means if you are performing a two sample test).

So, now here I have given the details what I just explained you that x and y are the data vectors, alternatives is the specification for the alternative hypothesis. This can be two sided greater or less a two-sided means $H_1$: $\mu \neq \mu_0$, greater means $H_1$: $\mu > \mu_0$ and s means $H_1 : \mu < \mu_0$ and μ here is the value of $\mu_0$. That is as simple as that.

(Refer Slide Time: 14:55)



**Hypothesis Testing for μ (σ unknown): *t* - Test in R**

Arguments

`paired` a logical indicating whether you want a paired t-test.

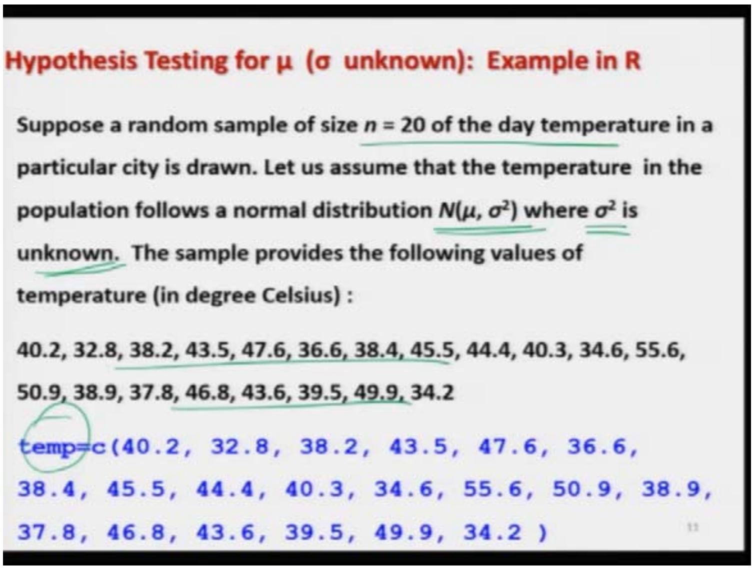
`var.equal` a logical variable indicating whether to treat the two variances as being equal. If `TRUE` then the pooled variance is used to estimate the variance otherwise the Welch (or Satterthwaite) approximation to the degrees of freedom is used.

$1-\alpha$

`conf.level` confidence level of the interval.

And simply I have given you here the option for paired, that is for the two sample paired t test and variance dot equal, v-a-r dot equal that means the variances are equal or not and confidence level this is the value of 1 − α.

(Refer Slide Time: 15:12)



**Hypothesis Testing for μ (σ unknown): Example in R**

Suppose a random sample of size *n* = 20 of the day temperature in a particular city is drawn. Let us assume that the temperature in the population follows a normal distribution *N*($\mu$, $\sigma^2$) where $\sigma^2$ is unknown. The sample provides the following values of temperature (in degree Celsius) :

40.2, 32.8, 38.2, 43.5, 47.6, 36.6, 38.4, 45.5, 44.4, 40.3, 34.6, 55.6, 50.9, 38.9, 37.8, 46.8, 43.6, 39.5, 49.9, 34.2

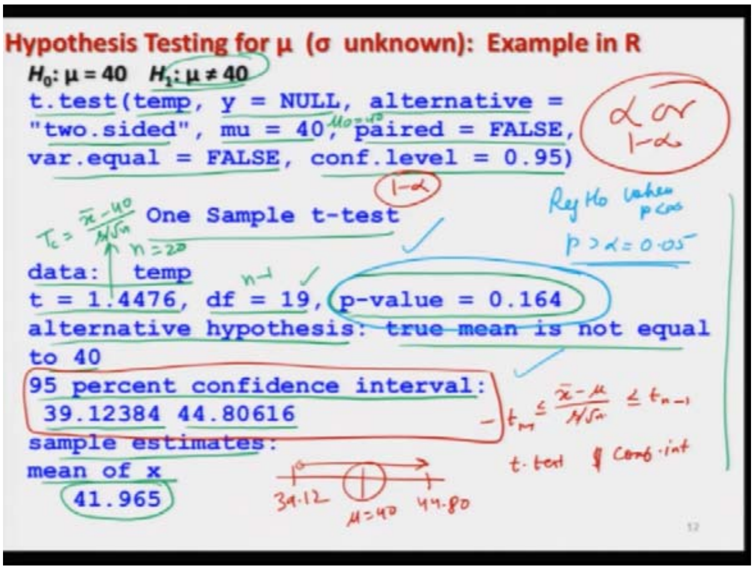
```
temp=c(40.2, 32.8, 38.2, 43.5, 47.6, 36.6,
38.4, 45.5, 44.4, 40.3, 34.6, 55.6, 50.9, 38.9,
37.8, 46.8, 43.6, 39.5, 49.9, 34.2 )
```

So, now using these options you can conduct any sort of test of hypothesis. So, let me try to take here the same example that I considered earlier. Well, you can ask me why I am considering the same example again and again and I am just modifying smaller things. My idea is very simple that I want to give you a feeling that what will happen when under what type of conditions what are the statistical changes that can occur in the data and the related outcome.

So, now you have considered here the same example in different situations. So, now you can compare the outcomes when for example $\sigma^2$ is known, $\sigma^2$ is unknown, you are trying to change the level of significance when you are trying to construct the confidence interval etc.

So, this will give you a complete analysis or you can consider that many times people ask for a case study, so now in this case study I have taken the same data set and I have employed all possible tools. So, now in this example we have the data of 20 temperature of the day means day temperature in a city and that is assumed to follow a normal μ $\sigma^2$ distribution where $\sigma^2$ is now unknown. Earlier we had taken it to be known. So, now this data is here like this which has been stored in a variable temp.

(Refer Slide Time: 16:37)



Hypothesis Testing for μ (σ unknown): Example in R

$H_0: \mu = 40 \quad H_1: \mu \neq 40$

```
t.test(temp, y = NULL, alternative =
"two.sided", mu = 40, paired = FALSE,
var.equal = FALSE, conf.level = 0.95)

        One Sample t-test

data:  temp
t = 1.4476, df = 19, p-value = 0.164
alternative hypothesis: true mean is not equal
to 40
95 percent confidence interval:
 39.12384 44.80616
sample estimates:
mean of x
   41.965
```

$\mu_0 = 40$
α or 1-α
1-α
$T_c = \frac{\bar{x}-40}{s/\sqrt{n}}$
n=20
n-1
Rej H0 when p<α
p > α = 0.05
$-t_{n-1} \le \frac{\bar{x}-\mu}{s/\sqrt{n}} \le t_{n-1}$
t-test & Conf. int
39.12
μ=40
44.80
12

So, now we try to use here the command on the R software and we try to solve the same example. So, suppose we are interested in testing the hypothesis $H_1$: μ ≠ 40 degree Celsius. So, now how to input this data? How to read the command? How to give the instructions to the command? And how to read the different outcomes of the software? That is what we have to understand here and here I am going to explain in detail and after that in all other cases I will simply follow the same rule.

So, you see here in order to test this hypothesis, first we take a call that which of the test we have to use so because here $\sigma^2$ is unknown to us, so we decide for the t test. So, the command will become here t dot test and now we try to give the input data which is here in the data vector temp that is temperature. then the second data vector is going to be null because we are considering here only the one sample t test.

Then the alternative has to be specified based on this $H_1$: $\mu \neq 40$ which is here 2 dot sided. Now, you have to specify here the value of μ which is here the value of μ naught which is equal to here 40. Now, you will see here paired is equal to FALSE because you are not considering here 2 sample and then variance dot equal to fall because you are again considering only the one sample test and confidence level that is conf dot level that is equal to 0.95. And if you try to see here, this is the outcome that you will get in the R console.

So, now you have to understand what this software is trying to explain you. So, the first line here is the name of the test which is one sample t test and then what data you have used, that is temp. Now, here it is giving the value of here t so t is the value of here $T_c$ which was $\bar{x} - \mu_0$ which is actually here 40 divided by s by root n. So, and here is 20, so now this value is computed for that $\sigma^2$ is estimated by $s^2$. It is replaced and then this quantity comes out to be 1.4476. Now, this is here df. Df means degrees of freedom so this is n – 1, so this is 20 – 1 which is 19.

Now, this is the value of here p that is p-value is coming out to be 0.164 that is computed by the software and then it is indicating that what is the alternative hypothesis that a true means is not equal to 40. So, looking at the software outcome also you can know that which of the hypothesis is being tested here. Then you see here it is also giving you the 95 percent confidence interval which is between 39.12384 to 40.80616 and sample estimate for the μ that is here arithmetic mean of x is coming out to be 41.965. Now, I ask you to think what is this thing.

Do you remember that when we had learned the confidence interval for mean when $\sigma^2$ is unknown? Then we had used the statistics $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ to calculate that this value is lying between $-t_{n-1}$ to $t_{n-1}$ and then we had computed our confidence interval and in that situation when I was trying to explain you how to obtain the confidence interval on the basis of R software then I had given you an example and this is the same example which I am considering here and I had told you that if you want to obtain the confidence interval, you have to use the command here like t dot test and then you have to give here a dollar command and then you have to write down here conf dot int and so on.

So, at that moment, if you try to see and if you try to recall I will explain you that at this moment I am trying to give you only the statement but why this statement is coming, how this statement is coming that I will be in the condition to explain you later on. So, this is the place where I can explain you that you can remember that in the first lecture on the test of hypothesis, we had considered that test of hypothesis and confidence interval estimation they are also inter-related and looking into the value of confidence interval you can also take a call whether your null hypothesis is going to be accepted or rejected.

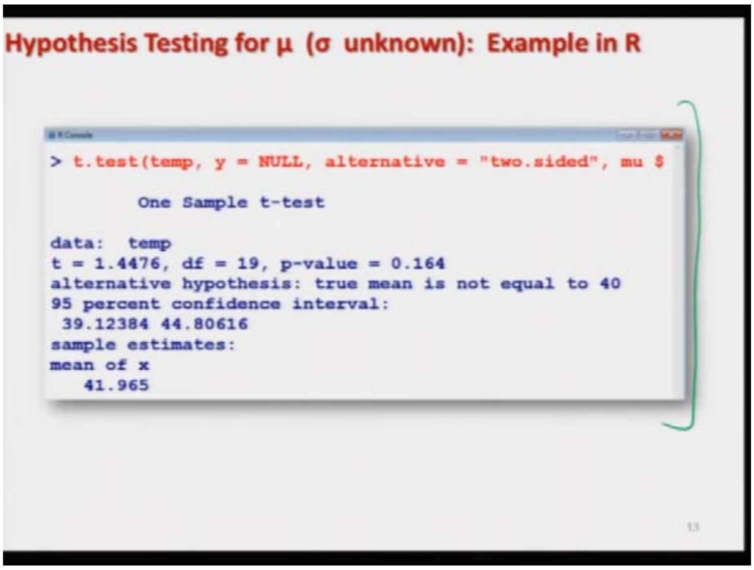
So, that is why in this software this is also giving you the value of confidence interval. So, since you are giving here conf dot level is equal to 0.95, so this is giving you here the 95 percent confidence interval. Now, you can see whether this value of $H_0 : \mu = 40$ is lying inside this interval or not. So, you can see here the lower limit of the confidence interval is around 39.12 and upper limit is here 44.80 and your this $H_0 : \mu = 40$ is saying that $\mu$ here is like 40. So, you can see here this value $H_0 : \mu = 40$ is going to be lie in this interval and when this value is lie in this interval, then we say that the $H_0$ is accepted.

So, in this case you can accept yes the temperature in the population is close to 40 degree and this is the same thing which you can observe here also, the p-value is coming out to be 0.164. So, p-value is greater than value of $\alpha$, $\alpha$ is equal to here 0.05. So, you can see here that the rule was reject $H_0$, when p is less than $\alpha$. So, now you can see here that p is greater than $\alpha$, so we are going to accept the hypothesis and the same conclusion that you can draw on the basis of p-value that is also drawn on the basis of confidence interval.

So, this is what I wanted to explain you. So, I hope I am clear here one thing you have to keep in mind that whenever you are trying to work in the software, you have to understand what you need to specify. Whether you need to specify the value of $\alpha$ level of significance or you need to specify $1 - \alpha$ which is the confidence coefficient. So, that is obtainable this information can be known only after looking into the instructions and help menu of the software for that test. So, for example in the case of this t test you have to specify the value of here 1 – α.

So, in case if you decide that α=0.05, then you have to specify here 1 – 0.05 which is 0.95. So, I hope I have made everything clear in this software outcome and I have explained you how are you going to take the outcome. Now, in the next cases this will go very simply straight forward.

(Refer Slide Time: 23:42)

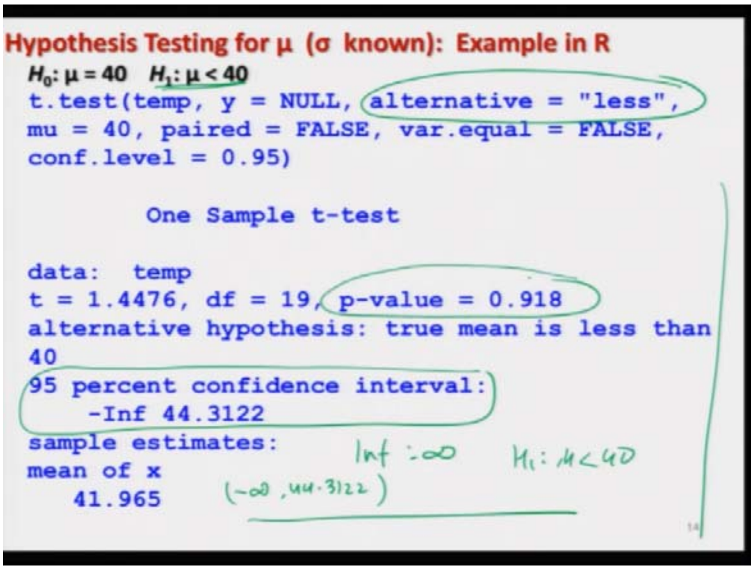


Hypothesis Testing for μ (σ unknown): Example in R

```
> t.test(temp, y = NULL, alternative = "two.sided", mu $

        One Sample t-test

data:  temp
t = 1.4476, df = 19, p-value = 0.164
alternative hypothesis: true mean is not equal to 40
95 percent confidence interval:
 39.12384 44.80616
sample estimates:
mean of x
   41.965
```

So, this is here the screenshot of the same test which I shown you. Anyway I will show you on the R console also.

(Refer Slide Time: 23:48)



Hypothesis Testing for μ (σ known): Example in R

$H_0: \mu = 40$ $H_1: \mu < 40$

```
t.test(temp, y = NULL, alternative = "less",
mu = 40, paired = FALSE, var.equal = FALSE,
conf.level = 0.95)

        One Sample t-test

data:  temp
t = 1.4476, df = 19, p-value = 0.918
alternative hypothesis: true mean is less than
40
95 percent confidence interval:
    -Inf 44.3122
sample estimates:
mean of x
   41.965
```

Inf :-∞ $H_1: \mu < 40$

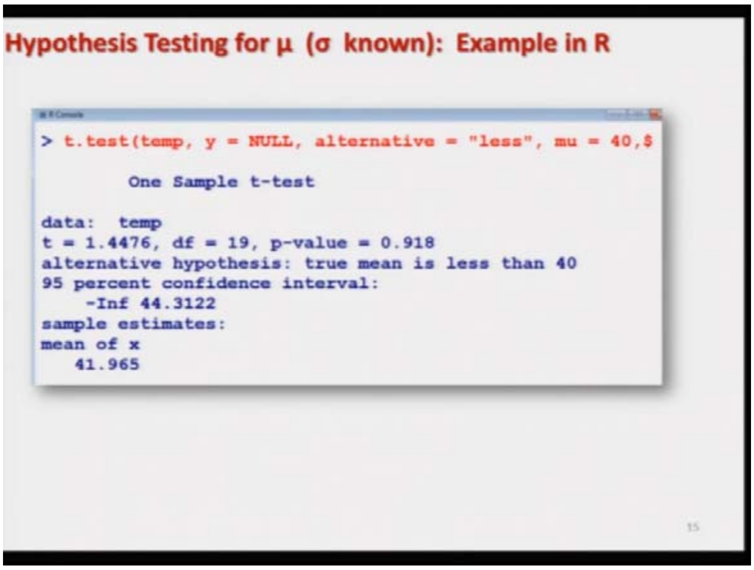
$(-\infty, 44.3122)$

14

Now, in case if you try to change your alternative that $H_1 : \mu < 40$ then the entire command everything remain the same only thing becomes here that alternative is going to be changed. Now, alternatively becomes say inside the double quotes you have to write less and then you have here the this outcome and this outcome can exactly be seen in the same way as we have done earlier. But you can see here that 95 percent confidence interval is coming out to be like here – infinity to 44.3122.

So, this is i-n-f means infinity, so this is the interval – infinity to 44.3122 and if you recall that in the case of have confidence interval estimation, we have done two sided confidence interval, one sided confidence interval. So, this is your here one-sided confidence interval and if you can recall, we had obtained that the value is going to be start from – infinity to the point or from the point to infinity. So, in this case when you are trying to consider the alternative hypothesis like $H_1 : \mu < 40$ , then you are going to have an interval like this one and again you can take the conclusion through the use of p-value or you can also take the same conclusion from the confidence interval also.

(Refer Slide Time: 25:05)



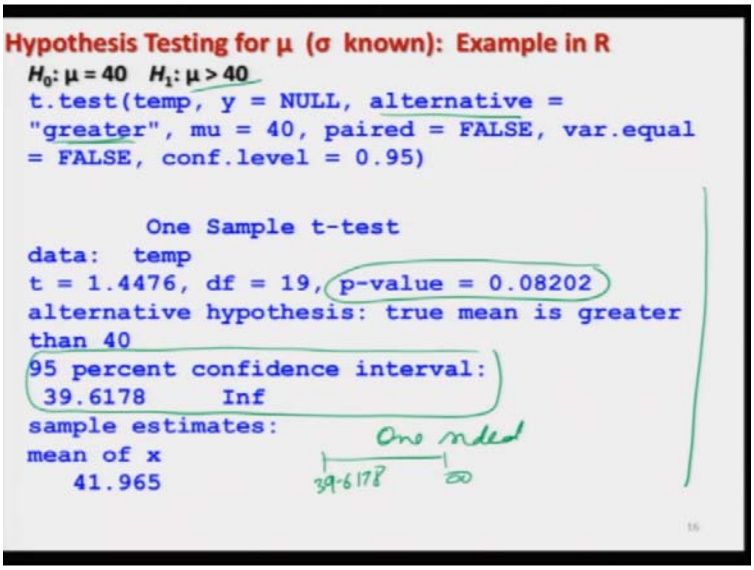
Hypothesis Testing for μ (σ known): Example in R

```
> t.test(temp, y = NULL, alternative = "less", mu = 40,$

        One Sample t-test

data:  temp
t = 1.4476, df = 19, p-value = 0.918
alternative hypothesis: true mean is less than 40
95 percent confidence interval:
    -Inf 44.3122
sample estimates:
mean of x
   41.965
```

15

And this is the screenshot, this I will try to show you on the R console.

(Refer Slide Time: 25:07)



Hypothesis Testing for μ (σ known): Example in R

$H_0: \mu = 40 \quad H_1: \mu > 40$

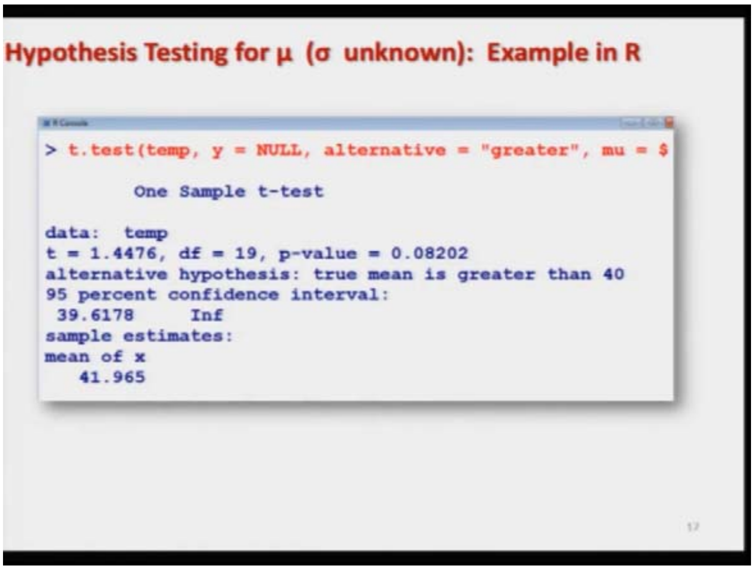
```
t.test(temp, y = NULL, alternative =
"greater", mu = 40, paired = FALSE, var.equal
= FALSE, conf.level = 0.95)

        One Sample t-test
data:  temp
t = 1.4476, df = 19, p-value = 0.08202
alternative hypothesis: true mean is greater
than 40
95 percent confidence interval:
 39.6178      Inf
sample estimates:
mean of x
   41.965
```

16

And this is the last hypothesis that $H_1 : \mu > 40$ . So, in this case you simply have to use the same command alternative is going to be now greater. So, now if you try to execute this command this is here the outcome and you can see here the outcome is exactly the same as earlier, only the p-value is going to be changed and you can see here the confidence interval is also now changing. The confidence interval is now becoming from 39.6178 to infinity.

So, this is a one sided confidence interval and that is the same thing that you learnt when you did the confidence interval estimation. So, now you can see here that everything is now clear, how are we going to consider the test of hypothesis, how are we going to implement it, how are we going to interpret it and the interpretation can be made on the basis of p-value or on the basis of confidence intervals. You will see that in some of the test of hypothesis in the outcome sometimes you will get the confidence interval, sometimes you will not get the outcome containing the confidence intervals. So, but now you know whatever you want to do you can do it very easily.

(Refer Slide Time: 26:19)

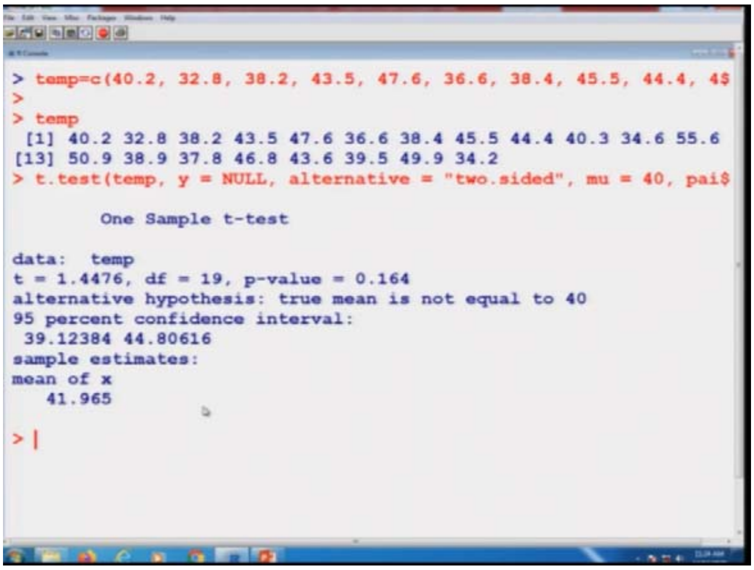


Hypothesis Testing for μ (σ unknown): Example in R

```
> t.test(temp, y = NULL, alternative = "greater", mu = $

        One Sample t-test

data:  temp
t = 1.4476, df = 19, p-value = 0.08202
alternative hypothesis: true mean is greater than 40
95 percent confidence interval:
 39.6178      Inf
sample estimates:
mean of x
   41.965
```

And this is here the screenshot of the same thing.

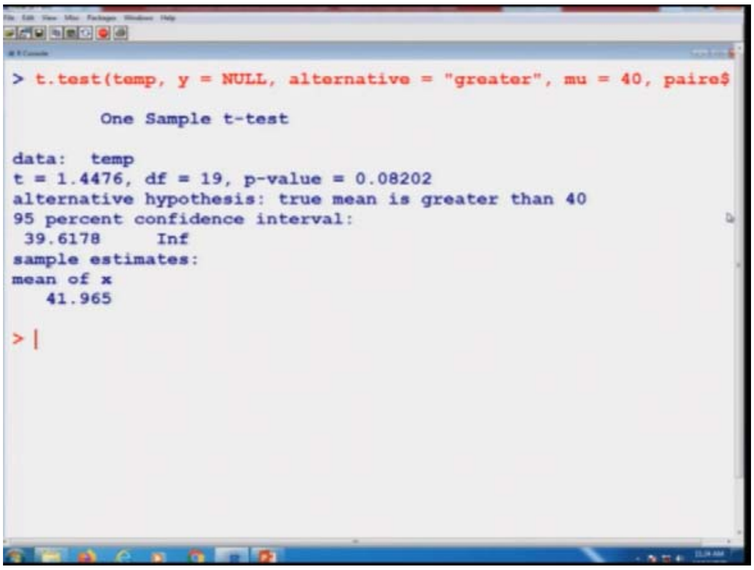
(Refer Slide Time: 26:22)



```
> temp=c(40.2, 32.8, 38.2, 43.5, 47.6, 36.6, 38.4, 45.5, 44.4, 4$
>
> temp
 [1] 40.2 32.8 38.2 43.5 47.6 36.6 38.4 45.5 44.4 40.3 34.6 55.6
[13] 50.9 38.9 37.8 46.8 43.6 39.5 49.9 34.2
> t.test(temp, y = NULL, alternative = "two.sided", mu = 40, pai$

        One Sample t-test

data:  temp
t = 1.4476, df = 19, p-value = 0.164
alternative hypothesis: true mean is not equal to 40
95 percent confidence interval:
 39.12384 44.80616
sample estimates:
mean of x
   41.965

> |
```

But now let me try to first come on the R console and try to show you that whether these things are really working or not. So, first let me try to copy here the data, here temp temperature. So, you can see here this is the data here temp and now you have to simply use the command here for the t test. So, for the two-sided t test you can see here this is here the outcome and in case if you want to make it here unsided test say here less than type. So, this is the alternative l-e-s-s has to be changed and you can see here this is here the outcome for the $H_0$ μ less than 40.

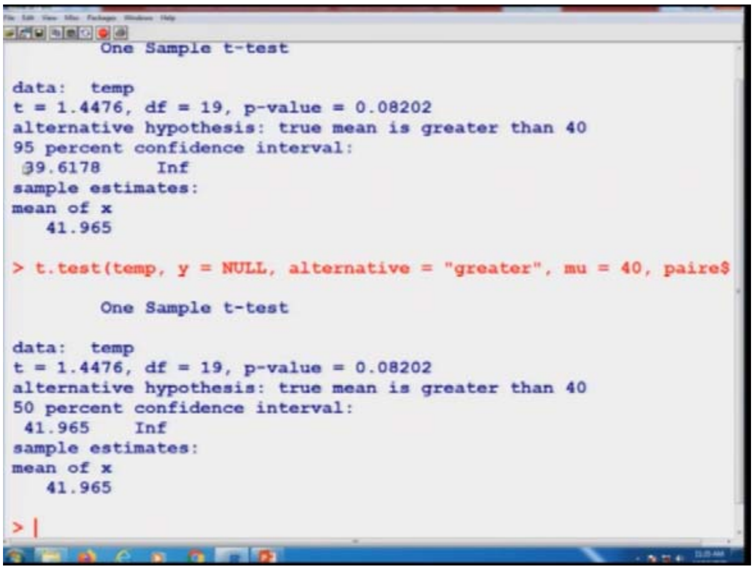
(Refer Slide Time: 27:10)



```
> t.test(temp, y = NULL, alternative = "greater", mu = 40, paire$

        One Sample t-test

data:  temp
t = 1.4476, df = 19, p-value = 0.08202
alternative hypothesis: true mean is greater than 40
95 percent confidence interval:
 39.6178     Inf
sample estimates:
mean of x
   41.965

> |
```

And in case if you want to do it for the greater than then you have to simply give here the command in which you have to change the alternative to be only here greater. So, you can see here this is greater and the outcome here is like this. So, you can see here that everything is going to be changed.

(Refer Slide Time: 27:26)



```
        One Sample t-test

data:  temp
t = 1.4476, df = 19, p-value = 0.08202
alternative hypothesis: true mean is greater than 40
95 percent confidence interval:
 39.6178      Inf
sample estimates:
mean of x
   41.965

> t.test(temp, y = NULL, alternative = "greater", mu = 40, paire$

        One Sample t-test

data:  temp
t = 1.4476, df = 19, p-value = 0.08202
alternative hypothesis: true mean is greater than 40
50 percent confidence interval:
 41.965     Inf
sample estimates:
mean of x
   41.965

> |
```

And in case if you try to change here the suppose, if I try to change here the confidence level to suppose here 50 that I am saying that okay the Type One error is going to be 50 percent well it is not acceptable but still I want to show you something that is why I am doing it here. You can see here this confidence interval, this is in the first case when your confidence level is 95 percent it is 39.6178 to infinity but if you try to increase it to 50 percent this is also changing from 41.965 to infinity. So, and similarly, if you can see here for say α equal to 0.05 and 0.5 the p-values are going to remain the same.

Why? because p-values are sample dependent. In this case if you try to look the values from the table means for example if you are trying to do this calculation using the calculated values and tabulated values from the t table, then if you try to change the value of α this tabulated values are going to change and then your decision may change but that is the advantage actually of using the p-value and once you are trying to use the p-value, these are the sample dependent value and the software is computing them.

So, that is why whatever you $\alpha$ you want to choose this p-value is not going to change but your decision will change because your rule was reject $H_0$ when p-value is less than $\alpha$. So, either you are trying to compare the p value with 5 percent or 50 percent that is your choice. So, this is how we are going to conduct the test of hypothesis.

One point I would like to emphasize towards the end of this lecture that many times you will read a different type of literature that there is a discussion whether this p-value should be used or p-values are dependable etc. etc. So, definitely I am not going into that discussion, my job is to make you learn and I have given you both the approaches. If you are trying to compute the calculated value of t statistics, you are trying to find out the tabulated value of the t statistics from the probabilities of t table or you are trying to use the software to compute the p-values you should know both the things definitely.

The first approach has no issues everybody believes on it that if you try to find out the value of t statistics on the sample of data and then you try to find out the tabulated value from the t table and if you try to take a decision means everybody will agree without any discussion without any confusion.

But I wanted that you should know this thing that because in data sciences you are trying to do all those things in an automated way and these types of things are very popular in data science and particularly when you are trying to do the different types of statistical modeling, this test of hypothesis are the backbone of the statistical modeling.

Without them I can guarantee you cannot do any type of modeling including the linear regression modeling and that was the reason I had chosen the topic of this course to be essentials of data science. After this, whether you want to do data science or not that is up to you but without them you cannot do it that is the guarantee.

So, I will stop now in this lecture, I will try to request you that you please try to look into assignment your books try to get some more example and then I will try to take up the issue of testing of say mean in some other condition in the next lecture. So, I will try to see you in the next lecture and till then good bye.